Extracting Chemical Reactions from Biological Literature

Jeffrey Tsui

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-109
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-109.html

May 16, 2014

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **16 MAY 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE
Extracting Chemical Reactions from Biological Literature

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California at Berkeley,Electrical Engineering and Computer Sciences,Berkeley,CA,94720

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Synthetic biologists must comb through vast amounts of academic literature to design biological systems. The majority of this data is unstructured and difficult to query because they are manually annotated. Existing databases such as PubMed already contain over 20 million citations and are growing at a rate of 500,000 new citations every year. Our solution is to automatically extract chemical reactions from biological text and canonicalize them so that they can be easily indexed and queried. This paper describes a natural language processing system that generates patterns from labeled training data and uses them to extract chemical reactions from PubMed. To train and validate our system, we create a dataset using BRENDA, the BRaunschweig ENzyme DAtabase, with 4387 labeled sentences. Our system achieves a recall of 0.82 and a precision of 0.88 via cross validation. On a selection of 600,000 PubMed abstracts, our system extracts almost 20% of existing reactions in BRENDA as well as many that are novel.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**18**	

Extracting Chemical Reactions from Biological Literature

Jeffrey Tsui
Master of Science in Computer Science
University of California, Berkeley
Advisor: Ras Bodik

Abstract

Synthetic biologists must comb through vast amounts of academic literature to design biological systems. The majority of this data is unstructured and difficult to query because they are manually annotated. Existing databases such as PubMed already contain over 20 million citations and are growing at a rate of 500,000 new citations every year. Our solution is to automatically extract chemical reactions from biological text and canonicalize them so that they can be easily indexed and queried. This paper describes a natural language processing system that generates patterns from labeled training data and uses them to extract chemical reactions from PubMed. To train and validate our system, we create a dataset using BRENDA, the BRaunschweig ENzyme DAtabase, with 4387 labeled sentences. Our system achieves a recall of 0.82 and a precision of 0.88 via cross validation. On a selection of 600,000 PubMed abstracts, our system extracts almost 20% of existing reactions in BRENDA as well as many that are novel.

Table of Contents

1. Introduction

Synthetic biologists are interested in developing novel biological systems that produce compounds such as biofuels and pharmaceutical drugs. Researchers build these systems by combining dozens of enzymes and metabolic pathways, which are constantly being discovered. While most important findings are published, they are available in the form of unstructured text and are difficult for researchers to discover and incorporate into new research. Biologists as well as other researchers need the ability to perform comprehensive queries over academic literature to leverage findings from new research.

One of the most widely used resources in biology is PubMed, a database consisting of roughly 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. Researchers may search through PubMed by forming queries using MeSH terms. MeSH (or Medical Subject Headings) is the National Library of Medicine's controlled vocabulary for indexing publications and chemicals and enzymes. MeSH tags are assigned manually[1], a slow process that results in many incompletely labeled publications.

We describe a sample workflow using PubMed to illustrate incomplete annotation. A synthetic biologist is interested in searching for a novel enzymatic pathway to produce D-glucaric acid, one of twelve high-value bio-based chemical building blocks named by the U.S. Department of Energy in 2004.[2] She creates the following query and retrieves 805 results from PubMed’s website.

```
("glucaric acid"[MeSH Terms] OR ("glucaric"[All Fields] AND "acid"[All Fields]) OR "glucaric acid"[All
Fields])
```

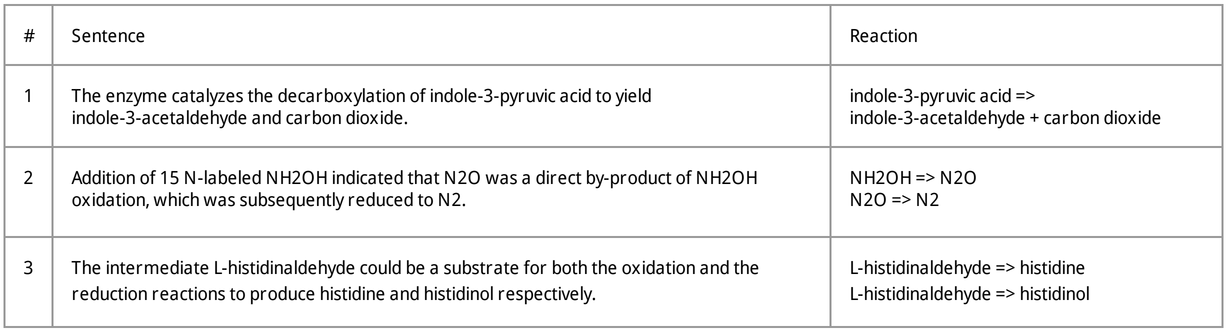
Our biologist must read through all 805 results manually, but even after reading through all of them the results are not comprehensive. The following paper[3] identified by our pattern matching system is not returned in the search results but describes a potential reaction involving glucaric acid that may be of interest. Below is an excerpt from its abstract:

```
The membrane-bound quinoprotein glucose dehydrogenase (mGDH) in Escherichia coli contains
pyrroloquinoline quinone (PQQ) and participates in the direct oxidation of D-glucose to D-gluconate by
transferring electrons to ubiquinone (UQ).
```

Researchers also have access to other databases such as BRENDA and KEGG which are more structured than PubMed, providing additional information about enzymes and other molecular interactions. However, both databases are manually curated and have the same weakness as PubMed. While biologists have a large amount of data at their disposal, much of the data is in the form of unstructured text and not easily accessible. As the rate of growth in the number of publications increases, manual tagging will require more time and resources.[4]

We solve this problem by building a natural language processing system to automatically extract chemical reactions from PubMed. The system uses labeled training data to generate a set of patterns that are used to extract reactions. The resulting reactions are canonicalized by converting reactants to standard InChI notation, which can augment existing databases and search tools. We focus on extracting reactions from individual sentences in PubMed abstracts, leaving the problem of extracting reactions that are described across multiple sentences as future work. Sentences may describe multiple reactions, each consisting of two or more reactants. As shown in table 1, the language used to describe reactions is surprisingly variable and non formulaic.

[1] http://www.nlm.nih.gov/bsd/disted/meshtutorial/principlesofmedlinesubjectindexing/theindexingprocess/
[2] http://www1.eere.energy.gov/bioenergy/pdfs/35523.pdf
[3] http://www.ncbi.nlm.nih.gov/pubmed/?term=16216080
[4] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909426/



#	Sentence	Reaction
1	The enzyme catalyzes the decarboxylation of indole-3-pyruvic acid to yield indole-3-acetaldehyde and carbon dioxide.	indole-3-pyruvic acid => indole-3-acetaldehyde + carbon dioxide

2	Addition of 15 N-labeled NH2OH indicated that N2O was a direct by-product of NH2OH oxidation, which was subsequently reduced to N2.	NH2OH => N2O
		N2O => N2

3	The intermediate L-histidinaldehyde could be a substrate for both the oxidation and the reduction reactions to produce histidine and histidinol respectively.	L-histidinaldehyde => histidine
		L-histidinaldehyde => histidinol

Table 1: Example sentences from PubMed abstracts.

In the rest of the paper, we first discuss related works (Section 2), followed by our pattern representation and the process of extracting a reaction from a sentence given a pattern (Section 3). Next we describe the process of creating a training set using reactions in the BRENDA database (Section 4) which is used to automatically generate patterns (Section 5). We evaluate the patterns generated by our system and present results in Section 6. Finally, we summarize limitations in our approach and outline future works (Section 7) before concluding.

2. Related Works

Semantic relationship extraction is the task of extracting relationships between tagged entities in text. The problem can be framed in the following way: given a document with entities labeled e_1, e_2, ... e_n, extract a set of relationships between any subset of entities. Traditionally the types of relationships being extracted are simple binary relations between two entities. These relations include location (Cal is located in Berkeley) and father-son (George HW Bush is the father of George W Bush). Relationship extraction has been framed in two ways: binary classification and bootstrapping pattern-based approaches.

The classification approach assigns +1, -1 (has a relation or doesn't) between every pair of entities (for binary relations), given a set of features extracted from the sentence. The features are often selected heuristically using expert knowledge, making it difficult to define an optimal set of features that work well across multiple domains. Once features are selected, they can be used in SVMs and other standard classifiers. One general feature that is commonly used is the string kernel, which is described by Lodhi et al. to measure the number of common subsequences between two strings. A function K(x,y) computes the similarity between two strings x and y, which can be directly incorporated as the similarity metric in perceptron classifiers. The subsequences in the kernels can be individual characters, words, or parse trees.

DIPRE (Brin) is one of the first pattern based approaches for relationship extraction. It was originally used to extract the author relationship: i.e. The author of *The Adventures of Sherlock Holmes* is Arthur Conan Doyle. DIPRE generates patterns consisting of a list of six elements: order, author, book, prefix, suffix, and middle. The prefix, suffix, and middle components are fixed length sequences of characters surrounding mentions of the author and book. We adopt an approach based on the Rapier system (Califf), which describes a bottom up approach to generate pattern matching rules for binary relations. Instead of sequences of characters used in DIPRE, Rapier uses lists of tokens for the prefixes and suffixes surrounding entities to be extracted. The tokens consist of syntactic elements such as a word's part of speech tag. There is no limit on the length of each list, and patterns are generated iteratively with increasing lengths until an optimal pattern set is found. Optimality is approximated when the score improvement between subsequent patterns sets fall below some threshold.

Relationship extraction has also been applied in the biology domain. Most relationships extracted are related to genes and proteins, using domain specific hand-written patterns. PASTA (Gaizauskaset et al.) uses type and POS tagging with manually created templates to extract relationships between amino acid residues and their protein function. Pustejovsky et al. uses a rule-based system to extract entities in the inhibit relation (x inhibits metalloproteinases). A useful tool for relationship extraction in biology is named entity recognition (NER). NER in biological text has been quite successful in extracting chemical names with F-scores between 75-85 percent. We use existing NER systems to preprocess abstracts before the pattern generation and extraction processes.

While relationship extraction has been studied in the past, the problem of extracting chemical reactions from biological literature has not been solved. Extracting chemical reactions is more difficult than the binary relations described above since they often involve more than two reactants. We adopt a pattern based approach over a classification approach. Extracting an optimal set of features relevant to chemical reactions is difficult and requires a significant amount of trial and error as well as domain expertise. Hand-built rules have been successful in the bioscience domain giving us confidence that automatically generating patterns similar to these rules will also be successful. Pattern based approaches also have the advantage that the patterns generated by the system are readable making analysis more transparent.

3. Extracting Reactions using Patterns

In this section we describe our pattern representation and outline the pattern matching process to extract a reaction. Figure 1 shows an overview of the extraction process starting with a sentence from an abstract. The sentence is first parsed grammatically into a syntax tree and labeled semantically with chemical and enzyme tags. Patterns match the syntactic and semantic labels to identify and extract the reactants in the sentence.

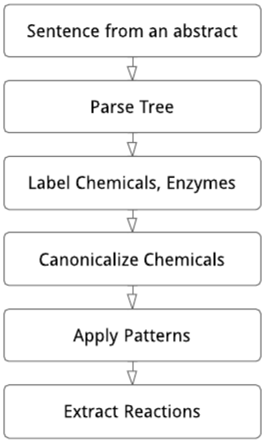


Figure 1: Process for extracting a chemical reaction from a sentence.

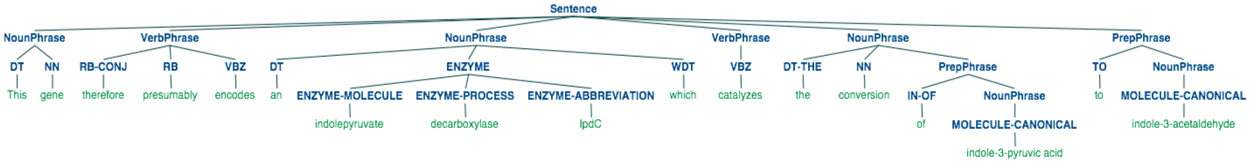


Figure 2: Example parse tree with syntactic and semantic (chemical and enzyme) labels.

3.1 Pattern Representation

Patterns are a set of constraints that limit matches to sentences that describe reactions. We select a pattern representation that is loose enough so that a single pattern can match multiple similar sentences. We consider top-down and bottom-up pattern representations. A top-down pattern keeps the hierarchical structure of the parse tree. In figure 2, the three rightmost subtrees (verb phrase, noun phrase, and prepositional phrase) are relevant to the reaction. An extremely general top down pattern may match any sentence that contains these phrases, allowing any variation of tokens in each subtree. This pattern would match a large number of sentences but would have many false positives. Including deeper portions of the tree would add more constraints and reduce the number of false positive matches. Some threshold is necessary to limit their complexity. A pattern requiring multiple nested subtrees to match is likely to be overly-specific and not useful.

Bottom-up patterns discard the tree hierarchy and have a flat structure. They match a sequence of tokens in a sentence (the leaves of the parse tree), and could be made more general by matching one of the token's tags. In figure 2, the leaves of the three rightmost subtrees may yield the following pattern *"catalyzes the conversion of <substrate> to <product>"*. A pattern that requires a sentence to match all of these tokens would be overly specific. Instead of exact tokens, the pattern can match each token's POS tag: *"VBZ DT-THE NN IN-OF <substrate> TO <product>"*. The optimal pattern may be some combination of specific and general elements. We adopt bottom-up patterns to emphasize pattern precision. By starting out with the most restrictive pattern, we can improve the generality of the pattern incrementally while limiting the number of false positives

Our system uses patterns made up of an ordered sequence of elements. There are two types of pattern elements: constraints that are required to match and holes that are used to mark a token for extraction, either as a substrate or product. Constraints can be stems, POS or semantic tags, or wildcards. A stem constraint requires the token's root to match the element. The stem constraint "convert", for example, is satisfied by any of the following tokens: "interconversion, "conversion", and "converted". POS and semantic tag constraints are more flexible; any token with the same part of speech or semantic label (enzyme or chemical) can match. Wildcards are the least restrictive constraint; any single token will match. Each pattern has a minimum of two holes: one for a substrate and one for a product.

3.2 Extraction Process

The extraction process matches a pattern on all consecutive sequences in the sentence. A successful match is found when each element in the pattern is matched by a token in the sentence. Tokens matching the holes are extracted as either substrates or products. Figure 3 shows an example of a pattern with 9 elements and a sentence it matches.

#	1	2	3	4	5	6	7	8	9
Type	stem	wildcard	hole	POS	hole	POS	hole	POS	hole
Element	convert	*	substrate	CC	substrate	TO	product	CC	product
Matching Token	conversion	of	tetrahydrodipicolinate	and	succinyl-CoA	to	L-2-(succinylamino)-6-oxopimelate	and	CoA

Figure 3: Example pattern that matches the following sentence - the matching tokens are numbered: Tetrahydrodipicolinate N-succinyltransferase THDP catalyzes the conversion[1] of[2] tetrahydrodipicolinate[3] and[4] succinyl-CoA[5] to[6] L-2-(succinylamino)-6-oxopimelate[7] and[8] CoA[9].

Long sentences may describe a substrate at the beginning, irrelevant details in between, and the product at the end. A single-sequence pattern will require tokens for all words in between the substrate and product which are irrelevant to the reaction. To account for cases like this, patterns may contain multiple sequences of tokens. A pattern with multiple sequences is matched separately and sequentially. If all sequences are matched, the extracted reactants from each sequence are combined into a single reaction.[5] For the long sentence described above, a two-sequence pattern would concisely describe the reaction, one to match the substrate and another to match the product.

3.2 Sentence Parsing

Sentences need to be converted into their parse trees for pattern matching. There are several existing named entity recognition (NER) tools and part of speech (POS) taggers trained on biochemistry text. We use ChemicalTagger, an open-source tool developed by the University of Cambridge's Center for Molecular Science Informatics, which performs both POS tagging and chemical entity extraction.

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner (a, an)
IN	Preposition - IN-OF, IN-INTO, IN-FROM
JJ	Adjective
NN	Noun - NN-SYNTHESIZE, NN-YIELD, NN-EXTRACT, NN-CHEMENTITY
PRP	Personal pronoun
TO	to
VB	Verb - VB-SYNTHESIZE, VB-YIELD
MOLECULE	MOLECULE-CANONICAL: molecules that are canonicalized to an InChI.
ENZYME	Ex. geranylgeranyl diphosphate phosphatase is labeled: ENZYME-ABBREVIATION ENZYME-MOLECULE: geranylgeranyl diphosphate ENZYME-PROCESS: phosphatase
OSCAR	Chemical entities labeled using OSCAR OSCAR-RN: reaction name (i.e. dephosphorylation) OSCAR-ASE: generic enzyme names (i.e. Reductase)

Table 2: Common part of speech tags used in patterns. The last two semantic tags "MOLECULE-CANONICAL" and "ENZYME" tags are labeled by our system while ChemicalTagger assigns the other tags.

We add two additional semantic labels to identify enzymes and organic chemicals. Sentences describing reactions often mention enzymes participating in the reaction. We label them so they can be matched by a pattern element. ChemicalTagger often labels enzymes as chemicals. We assign enzyme labels when chemicals end in "-ase". We further categorize enzymes into three components: abbreviation, process, and chemical. Geranylgeranyl diphosphate phosphatase is labeled with process "phosphatase" and chemical "geranylgeranyl diphosphate".

The second addition is chemical canonicalization. Many chemicals tagged by ChemicalTagger are inorganic

[5] Each pattern sequence can yield a set of extractions.Aa pattern with two sequences s1 and s2 can be applied to a sentence that yields matches {m1, m2} for s1 and {m3} for s2. The results are combined to form two reactions: m1 => m3 and m2 => m3.

compounds such as prolactin or generic compounds referred to by abbreviations such as RP1. These compounds can not be canonicalized to a chemical structure and are not useful to synthetic biologists. We use the Chemical Identifier Resolver (CIR) API[6] to convert chemical names to InChIs (IUPAC International Chemical Identifiers), a universal string encoding a chemical's structural information. Compounds that can be canonicalized are labeled "MOLECULE-CANONICAL" and patterns are restricted to extracting tokens with these tags.

4. Creating a Training Set

In this section, we describe the process of creating a training set consisting of sentences labeled with chemical reactions. We did not find such a dataset that already exists, so we create our own. Manually generating a large dataset requires a significant manual effort and domain expertise. Instead, we utilize BRENDA (the BRaunschweig ENzyme DAtabase) which contains roughly 50,000 reactions labeled with 31,236 PubMed papers that reference them. We use this data to seed the generation of our training set. The final training set we create consists of 4387 sentences, 2757 of which describe at least one reaction. In total there are 2996 reactions described in the training set.

Total Sentences	4387
Sentences with at least 1 reaction	2757
Sentences with no reactions	1630
Total reactions	2996

Table 3: Generated training set size.

4.1 Reaction Labeling

We use two methods to label the training set. The first searches for mention of the BRENDA reactant names in each sentence. If a sentence contains at least one substrate name and at least one product name, then the sentence is tagged with that reaction. We do not require all the reactants to be described in each sentence since papers often describe only the principal chemicals involved in the reaction. Abstracts often refer to the same chemical using many different names. For example, a search for butanol on PubChem, the chemical database maintained by NCBI, shows 345 alternative names such as n-butanol, 1-butanol, Butylic alcohol, and Butan-1-ol. We record synonyms for each chemical using PubChem and similar databases. However, the synonyms are not exhaustive so matching purely based on names misses many labels.

Due to the incompleteness of name matching, the second way we label sentences is using InChIs. We convert each chemical in a sentence to its InChI representation and further canonicalize InChIs by removing stereochemistry variations. Similar to name matching, if a sentence contains at least one substrate InChI and at least one product InChI, then the sentence is tagged with the reaction. The labels for the two methods of matching are merged and duplicates are removed, yielding the final training set. An overview of the training set generation process is shown in figure 4.

[6] http://cactus.nci.nih.gov/chemical/structure

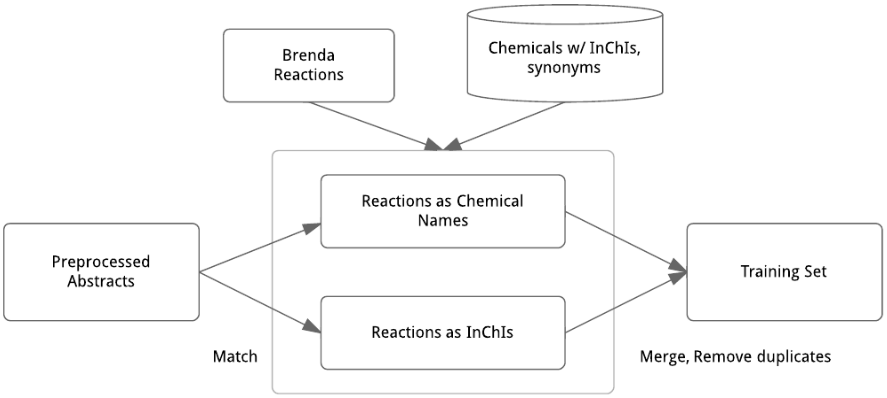


Figure 4: Training set generation process

There is a surprisingly large amount of gray area when it comes to determining whether two chemicals are the same. Besides many synonyms, certain terms express a general class of chemicals satisfying a specific substructure. For example, dioleoylglycerol represents any molecule that contains 2 fatty acids attached to glycerol. Its canonicalization is not unique; it describes a generic template for chemical structures. For these cases, the Chemical Identifier Resolver returns a single InChI of the most common instance of the chemical. The issue here is that for dioleoylglycerol and two instances of dioleoylglycerol X and Y, may match one but not the other. In most cases, the name matching process usually successfully matches variants of this form: the instances X and Y usually have dioleoylglycerol listed as a synonym.

4.2 Assumptions and Limitations

The training set is generated based on the assumption that if a sentence contains at least one substrate and at least one product, then it is talking about a reaction between them. We observe several cases where this is not true; a sentence may contain the reactants, but not describe the reaction between them. We attempt to remove some of these manually.

A portion of the training set will be used for cross validation. To realistically reflect the sentences in PubMed, the training set is augmented with 1630 negative samples. These samples are sentences from PubMed abstracts that contain at least two chemicals and do not describe reactions.

Finally, we observe many sentences that are incompletely tagged: a sentence that describes a reaction A -> B + C may only be tagged with A -> B. This could be the result of imperfect reactant matching (as noted above) or imperfect chemical extraction using ChemicalTagger’s NER. To compensate for these deficiencies in our evaluation, we define a successful pattern match as one where a subset of the reactants are extracted. In the section describing future works, we outline other ways of improving the training set.

5. Generating Patterns

This section describes the process of generating patterns. First the set of all possible patterns is enumerated using all combinations of tokens surrounding labeled reactants. This set contains 42 million patterns. Then we filter the patterns down to a set of roughly 800,000 and score them to select an optimal pattern set of roughly 1000 patterns. An optimal pattern set has high precision and recall, which we define in this context.

5.1 Pattern Enumeration

We first enumerate all possible patterns over the 2757 labeled sentences in the training set, resulting in 42 million patterns. For each (sentence, reaction) tuple we generate a set of patterns using the combination of tokens surrounding each reactant. The iterative process constructs patterns using tokens that are increasingly farther away from the labeled reactants.

The process is illustrated in table 4. Initially there is a single pattern containing a list of sequences, one for each reactant in the sentence. Next, we construct patterns using all combination of tokens of distance 1 away from each reactant. We continue this process for all combinations of tokens of at most distance 3 away from reactants. The maximum distance is set to 3 observationally to limit the total number of patterns generated. Tokens farther away from reactants usually do not describe the reaction. Each token can be included in a pattern as one of 3 constraint elements: stem, part of speech, or wildcard. For succinctness, only the words are shown in the table.

An exponential number of patterns is generated from each sentence: 3^n patterns are generated where n is the number of tokens less than 3 words away from reactants and there are 3 token variations for each word. With r reactants in a sentence, we generate at most 3^{6r} patterns patterns. Fewer patterns are generated when a reactant occurs at the beginning or end of a sentence.

Distance	Words Considered	Pattern Sets Extracted
0	The enzyme catalyzes the decarboxylation of indole-3-pyruvic acid to yield indole-3-acetaldehyde.	[[substrate], [product]]
1	The enzyme catalyzes the decarboxylation of indole-3-pyruvic acid to yield indole-3-acetaldehyde.	[[of substrate], [product]] [[substrate to], [product]] [[substrate], [yield product]] [[of substrate to], [product]] [[of substrate], [yield product]] [[substrate to yield product]] [[of substrate to yield product]]
2	The enzyme catalyzes the decarboxylation of indole-3-pyruvic acid to yield indole-3-acetaldehyde.	[[decarboxylation of substrate], [product]] [[decarboxylation of substrate to], [product]] [[decarboxylation of substrate], [yield product]] [[decarboxylation of substrate to yield product]]

Table 4: Patterns extracted from a sentence while iteratively expanding outwards from reactants. The patterns are shown without expanding each token to its variants, which include the stem, POS tag, and wildcard.

5.2 Pattern Scoring

To select an optimal set of patterns, we score the patterns according to its precision and recall. We first define precision and recall in this context.

When a pattern is applied on a sentence there are a number of possibilities. The first case is that the pattern does not extract a chemical reaction. If the sentence does not contain a chemical reaction, then this is a true negative. Otherwise, the sentence describes a chemical reaction and this is a false negative. The second case is that the pattern extracts a chemical reaction. If the sentence does not contain a reaction, this is a false positive. If the sentence describes a reaction but the extracted reaction does not match, this is also a false positive. The result is a true positive only when the extracted reaction matches a labeled reaction.

true positive	pattern extracts a reaction that matches the label
false positive	case 1: pattern extracts a reaction but there is no label case 2: pattern extracts a reaction but does not match the label
false negative	pattern does not extract a reaction but there is a label
true negative	pattern does not extract a reaction and there is no label

Table 5: Definitions of true positive, false positive, false negative, and true negative for a pattern.

As mentioned previously, we only require a subset of reactants to match. An extracted reaction r with substrates s and products p successfully matches a labeled reaction r' with substrates s' and p' if s is a subset of s' and p is a subset of p'. We also match reverse reactions since many descriptions of reactions do not specify direction or are reversible.

Using this notion of a good match, we score each pattern with its precision and recall. Precision is defined as $\frac{true\ positives}{true\ postiives + false\ positives}$ and measures what fraction of reactions extracted are valid. Recall is defined as $\frac{true\ positives}{true\ positives + false\ negatives}$ and measures what fraction of labeled reactions are retrieved. A good pattern will have high precision and recall.

5.3 Heuristic Filtering

Each pattern takes roughly 0.5 seconds to score, making it infeasible to score the complete set of 42 million patterns without first reducing it to a more manageable size. We do this by removing unique patterns, that is patterns that are only generated once across the entire training set. If a pattern is unique, it will only match the sentence it was generated from, leading to a very low recall. These patterns are likely to be overly specific, overfitting to a single sentence. After this filtering step, we reduce the working set from 42 million to just 800,000 patterns.

5.4 Pattern Selection

To select an optimal pattern set we score each pattern and record each set of successful matches. A pattern p can be subsumed when there is another pattern p' that satisfies two conditions: (1) p' has higher precision than p and (2) the set of successful matches for p' is a superset of that of p. Subsumed patterns should never be selected in the optimal pattern set because p' can be selected with better precision while matching the same sentences. Roughly 10% of the patterns are subsumed. From the remaining set of patterns, we select a single pattern for each reaction in the training set. The selected pattern matches the reaction with precision above some parameter p and has the largest f-score. A pattern's f-score is the weighted harmonic mean of its precision and recall:

$F_\beta = 1 + \beta^2 * \frac{precision * recall}{\beta^2 * precision + recall}$, using $\beta = 0.1$ (chosen observationally).

```
Pattern Selection Process

FOR each reaction in the training set
    get all patterns that match the reaction
    remove patterns with precision < p
    compute f scores of remaining patterns
    select the pattern with the highest f score
    add this pattern to optimal pattern set
ENDFOR
```

The selection method balances precision and recall of the final pattern set. By attempting to select a pattern for each reaction in the training set we improve the overall recall. To improve precision, we only select patterns that have precision greater than p, a parameter we test experimentally. Table 5 gives a few examples of automatically generated patterns. The final pattern set has roughly 1000 patterns.

#	Sentence	Pattern
1	A complete initial velocity study in both reaction directions suggests that the enzyme catalyzes the conversion of acetyl CoA and L-serine to O-acetyl-L-serine (OAS) and coenzyme A (CoASH) by a ping pong bi bi kinetic mechanism.	convert, *, substrate, CC, substrate, to, product, CC, product
2	The enzyme catalyzes the decarboxylation of indole-3-pyruvic acid to yield indole-3-acetaldehyde and carbon dioxide.	reaction, *, substrate, to, VB-YIELD, product, CC, product
3	N-Acetylaspartylglutamate (NAAG) a prevalent peptide in the vertebrate nervous system may be hydrolyzed by extracellular peptidase activity to produce glutamate and N-acetylaspartate.	2 sequences a) substrate, DT, *, peptide b) TO, prod, product, CC, product

Table 6: Example of automatically generated patterns

6. Evaluation

In this section we first validate our system’s pattern representation by evaluating a manually written pattern set. Next, we evaluate the pattern sets that are automatically generated by our system using different parameters. Finally, we extract reactions from a selection of PubMed sentences and compare the canonicalized results with the BRENDA database.

Throughout the evaluation, we use the notion of the precision and recall for a pattern set, which is similar to that of a single pattern. The caveat is that a pattern set could have many extractions for a single sentence, some of which could be duplicates. We remove duplicate extractions before computing the overall precision and recall.

6.1 Pattern Structure

We first validate our pattern representation and show that a small number of patterns is effective at extracting chemical reactions. We manually create a set of hand-written patterns and apply it on the training set, yielding a recall of 0.26 and precision of 0.74. The results support the need to automatically generate a larger number of patterns to improve the overall recall.

The patterns are generated by noticing that many chemical reactions are described using a trigger word such as reduced, converted, oxidized, etc. and a transition word such as from, into, by, etc. We create a set of five basic pattern templates listed in the table below. We also add more flexibility in each pattern by allowing an optional wildcard in between each token.

We present a few example extractions using hand-written patterns in table 8. The false positive example shows a valid extracted reaction; however, it does not match the labeled reaction which has CTP as the product. This illustrates an example where the labeled reaction is incomplete - the correctly labeled reaction should be 2-c-methylerythritol 4-phosphate => 4-diphosphocytidyl-2-c-methylerythritol + CTP. Manually correcting the incompletely tagged reactions would reduce the number of false positive extractions.

#	Pattern	Transition	Trigger (stemmed)
1	[trigger] cmp [transition] cmp	{from, to, into, by, are}	{reduc, convert, produc, form, oxid, transform, bioconvert, synthes, react, interconvert}
2	cmp [trigger] [transition] cmp	{from, to, into, by, are, yield}	{convert, oxid, produc, interconvert}
3	cmp [trigger] cmp	n/a	{yield}
4	[trigger] cmp and cmp	n/a	{convert, interconvert}
5	cmp [transition] [trigger] cmp	{is, ar}	{produc, metabolit}

Table 7: List of all hand-written patterns.

	Sentence	Extracted Reaction	Labeled Reaction
True Positive	Phosphoserine aminotransferase (PSAT EC 2.6.1.52) a member of subgroup IV of the aminotransferases catalyses the conversion of 3-phosphohydroxypyruvate to l-phosphoserine.	3-phosphohydroxypyruvate => l-phosphoserine	3-phosphohydroxypyruvate => l-phosphoserine
False Positive	We show that an enzyme isolated from cell extract of Escherichia coli converts 2-C-methylerythritol 4-phosphate into 4-diphosphocytidyl-2-C-methylerythritol by reaction with CTP.	2-c-methylerythritol 4-phosphate => 4-diphosphocytidyl-2-c-methyler ythritol	4-diphosphocytidyl-2-c-methyl erythritol => CTP
False Negative	Cystathionine beta-synthase (CBS) a pyridoxal 5'-phosphate (PLP) dependent enzyme catalyzes the condensation of serine and homocysteine to form cystathionine.	serine + cystathionine => homocysteine	n/a
True Negative	Preoptic stimulation of progesterone-treated ASR (P-ASR) did not induce greater release of FSH than in control ASR ESR or P-ESR.	n/a	n/a

Table 8: Example extractions of hand-written patterns.

6.2 Generated Patterns

We evaluate several automatically generated pattern sets using 10-fold cross validation. The data set is randomly split into a training portion and a test portion and the patterns generated using the training portion are used to extract reactions from the test portion. We do 10 splits and average the precisions and recalls.

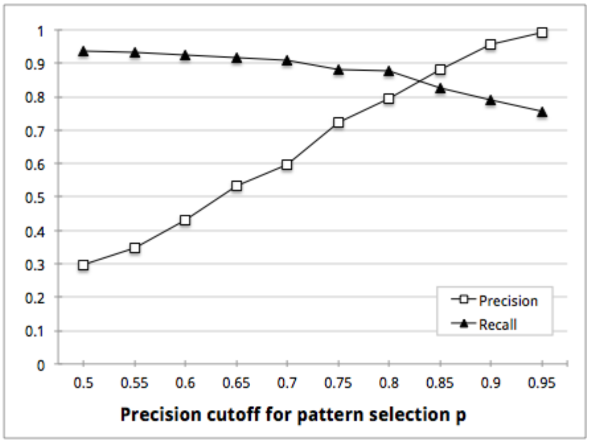


Figure 5: Precision and recall vs. p, the precision cutoff used for pattern selection. Using 10-fold cross validation with 25-75 split between test and training data, respectively.

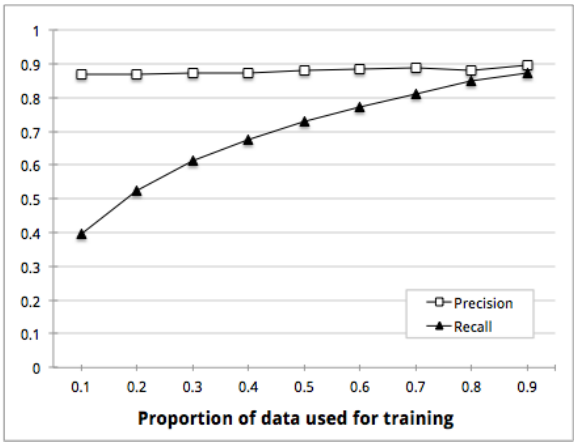


Figure 6: Precision and recall vs. split ratio used to split test and training in cross fold validation. Using 10-fold cross validation with p=0.85.

Figure 5 shows the precisions and recalls for 10 pattern sets generated with varying parameters of p. The parameter p is used in the pattern selection process to control the precision of the generated pattern set. The first datapoint is for a pattern set made up of patterns individually having precisions greater than 0.5. The precision of the overall pattern set is 0.31 with recall 0.93. Higher values of p will have a greater precision but lower recall. This is because some sentences may have no patterns with precision greater than p so we do not select a pattern for them. This illuminates an important aspect of our pattern-based approach. Patterns are designed to capture commonalities in phrases used to describe chemical reactions. Sentences that describe reactions using complicated, specific terms will not contribute meaningful patterns and they will not be matched by our system. We can vary p for different contexts depending on whether precision or recall is more important. We select a value of p=0.85 for further analysis since it strikes a good balance between precision and recall, with a precision of 0.88 and a recall of 0.82.

The recall is highly dependent on more training data. Figure 6 shows the precision and recall with fixed parameter p = 0.85, varying the proportion of data used for training during cross validation. The recall of the system increases significantly as the proportion of data used for training increases. A larger training set will generate patterns that match a greater variety of descriptions of chemical reactions.

Some example extractions are shown in table 9. The false positive example is due to incorrect chemical recognition. In the sentence, "lactic acid" is a chemical used as an adjective describing the bacteria and should not be labeled as a candidate for extraction. The false negative example demonstrates a limitation in the system. Although the system contains the pattern "*conversion of <substrate> into <product>*" (as seen in the true positive match), we did not match the false negative sentence which contains additional descriptions of the reactants "low activity" and "highly active". One way to make patterns more robust is to match phrases instead of individual tokens. Both "low activity" and "highly active" are adjectives that may be matched by phrase pattern elements such as "*conversion of JJ-phrase <substrate> into JJ-phrase <product>*". Before adopting this approach, the accuracy of syntax tree parsing must be improved. Currently, we have observed many POS tags to be incorrect, which would yield false positive extractions using phrase elements.

	Sentence	Extracted reaction	Labeled Reaction
True Positive	In the present study we report the purification and characterization of D-gluconate dehydratase from S solfataricus which catalyses the conversion of D-gluconate into 2-keto-3-deoxy-D-gluconate	d-gluconate => 2-keto-3-deoxy-d-gluconate	d-gluconate => 2-keto-3-deoxy-d-gluconate
False Positive	A study of the effects of histamine histidine and growth phase on histamine production by lactic acid bacteria isolated from wine is reported here.	lactic acid => histamine	n/a
False Negative	Human 17 beta-hydroxysteroid dehydrogenase 17-HSD type 1 catalyzes the conversion of the low activity estrogen estrone into highly active estradiol both in the gonads and in target tissues	n/a	estrone => estradiol
True Negative	MPOA-ECS of control ASR and ESR resulted in the release of LH and FSH	n/a	n/a

Table 9: Example extractions for automatically generated patterns.

We have noted several cases where the labels in the training set are incorrect or incomplete. We manually curate a subset of the training set and correct these errors to evaluate our patterns. The curation process involves selecting sentences from the training set and adding or removing incorrectly labeled reactants. The resulting 159 sentences describe reactions that can be easily extracted manually by a non-expert. We use the 159 sentences as a test set and generate patterns from the remainder of the training set, yielding a precision of 0.99 and recall of 0.90. The precision is much higher since we remove incorrect labels that are difficult for a non-expert to understand.

6.3 PubMed Extraction

We further evaluate our system by extracting reactions from a selection of PubMed that contains two or more chemicals and at least one enzyme. The enzyme restriction increases the proportion of sentences describing reactions because they are often described along with the catalyzing enzyme. We process all of PubMed and find 21,249,356 sentences with two or more chemicals. There are 1,060,007 sentences that mention an enzyme and 559,646 sentences that include an enzyme *and* have two or more chemicals. We generate patterns over the entire training set and extract reactions from the set of 559,646 sentences.

In total, we extract 47,829 unique reactions across 66,949 sentences. We compare the reactions extracted with the reactions in BRENDA and find that 8918 reactions already exist in BRENDA. This represents roughly 18% of the 50,000 reactions in BRENDA, showing that the recall of our system is quite high; roughly 20% of the sentences in the training set generated using BRENDA reactions have enzymes. The comparison uses canonicalized reactions where chemicals are represented as InChIs. A reaction is in BRENDA if it is a subset of an existing reaction.

We extract 38,911 reactions that are not in BRENDA. However, upon manual inspection many of these extractions are false positives and are not valid reactions. These false positives extracted by a subset of patterns that should have been removed in the selection process. Instead, the patterns were selected because we were unable to accurately compute their precision in the scoring process. Precision is calculated as $\frac{true\ positive\ matches}{all\ matches}$. Without enough training data, a poor pattern can have a high precision because few sentences yield a pattern match. The solution is to add more negative samples to the training set. With this addition we will observe more false positive matches, score poor patterns with lower precision, and remove them in the selection process. We describe further refinements of the training set in the following section.

7. Limitations and Future Work

7.1 Training Set Refinements

The main limitation of our pattern based system is its reliance on a large and variegated training set. The pattern set generated by the system can only extract reactions using the same phrasing as the ones in the training set. Furthermore, the pattern selection process uses the precision and recall of each pattern which will be inaccurate given a small training set. As we see in the PubMed evaluation, poor patterns may be inaccurately assigned a high precision if there are only a few sentences in the training set that it matches. In the evaluation section we note several false positive cases that are due to incompletely labeled sentences in the training set. Correcting these cases would also improve the pattern selection process. All of these refinements to the training set would improve the system but would require more manual effort.

7.2 Biological Filtering

Domain specific filtering could be used to reduce the number of false positive extractions. We have implemented a rudimentary filtering system that performs an atom-to-atom mapping on extracted chemical reactions. The process removes all chemical reactions that do not balance. The preliminary results of the filtering system show that the majority of *confirmed* reactions - roughly 70% do not pass the filtering step. This is because descriptions of partial reactions, although correct, will not balance. The reactions that *do* pass this filtering step are always correct. Domain specific filtering techniques like this can compensate for weaknesses in the natural language processing part of the system. They may be incorporated to other NLP systems in the biology domain to provide a greater level of confidence.

8. Conclusion

Currently, biological literature is manually annotated making it difficult to query for relevant chemical reactions. Although databases like PubMed contain a comprehensive collection of the literature, the data is in the form of unstructured text. In this paper, we describe a system that extracts chemical reactions from PubMed abstracts using patterns. Our system integrates existing tools to perform chemical canonicalization and syntax parsing. We describe our pattern representation and show that only a few hand-written patterns can successfully extract many reactions. This motivates the need for our system to automatically generate a large number of patterns to achieve high recall. We describe the process of creating a training set using BRENDA and the challenges in chemical canonicalization. Our system generates a set of roughly 1000 patterns with an overall recall of 0.82 and a precision of 0.88 via cross validation. Analysis on a selection of PubMed shows that the system has high recall, but further work is required to improve its precision. We note our system's reliance on a rich set of labeled training data and propose the use of domain filtering such as atom-to-atom mapping to further validate extractions.

Acknowledgments

This material is based upon work supported by the Defense Advanced Project Agency (DARPA) and Space and Naval Warfare Systems Center Pacific (SSC PAC) under Contract No. N66001-12-C-4204.

Special thanks to professors Ras Bodik and J. Christopher Anderson as well as Saurabh Srivastava for their advising throughout this project.

References

[1] Bach, Nguyen, and Sameer Badaskar. "A Review of Relation Extraction." : n. pag. Web.

[2] Barbara Rosario and Marti A. Hearst. 2004. Classifying semantic relations in bioscience texts. ACL 2004.

[3] Bundschus, M., Dejori, M., Stetter, M., Tresp, V. & Kriegel, H.-P. (2008). Extraction of semantic biomedical relations from text using conditional random fields.. BMC Bioinformatics, 9.

[4] "BRENDA The Comprehensive Enzyme Information System." . Braunschweig University of Technology. Web. 15 Dec 2013. <http://www.brenda-enzymes.org/>.

[5] Califf, Mary. "Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction."Journal of Machine Learning Research. (2003): n. page. Print. <http://www.cs.utexas.edu/~ai-lab/pubs/rapier-jmlr-03.pdf>.

[6] Chang, Antje. "BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009." Nucleic Acids Research. 37. (2009): n. page. Print. <http://www.ncbi.nlm.nih.gov/pubmed/18984617>.

[7] "Chemical Identifier Resolver." . NCI/CADD. Web. 15 Dec 2013. <http://cactus.nci.nih.gov/chemical/structure>.

[8] Cohen, A. M. & Hersh, W. R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6, 57-71.

[9] Freitag, Dayne. "Information Extraction from HTML: Application of a General Machine Learning Approach."1998. Print. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.8501>.

[10] Friedman, Carol. "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles." Bioinformatics. 17. (2001): S74-S82. Print. <http://bioinformatics.oxfordjournals.org/content/17/suppl_1/S74.full.pdf>.

[11] Gondy, Leroy. "A shallow parser based on closed-class words to capture relations in biomedical text."Journal of Biomedical Informatics. (2003): n. page. Print. <http://www.ncbi.nlm.nih.gov/pubmed/14615225>.

[12] Grego, Tiago. "Chemical Entity Recognition and Resolution to ChEBI." ISRN Bioinformatics. (2012): n. page. Print. <http://www.hindawi.com/isrn/bioinformatics/2012/619427/>.

[13] Hawizy, Lezan . "ChemicalTagger: A tool for semantic text-mining in chemistry." Journal of Cheminformatics. 3.17 (2011): n. page. Web. 15 Dec. 2013. <http://www.jcheminf.com/content/pdf/1758-2946-3-17.pdf>.

[14] Huffman, Scott. "Learning Information Extraction Patterns from Examples." n. page. Print. <http://www.cise.ufl.edu/~cgrant/projects/public/morpheus/files/learning_ir_patterns_from_examples.pdf>.

[15] Humphreys, Kevin. "Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures." Pacific Symposium on Biocomputing. (2000): 502-513. Print. <http://psb.stanford.edu/psb-online/proceedings/psb00/humphreys.pdf>.

[16] "Indigo Toolkit." . GGA Software Services. Web. 15 Dec 2013. <http://www.ggasoftware.com/opensource/indigo>.

[17] Miller, Scott. ALGORITHMS THAT LEARN TO EXTRACT INFORMATION BBN: DESCRIPTION OF THE SIFT SYSTEM AS USED FOR MUC-7. <http://acl.ldc.upenn.edu/muc7/M98-0009.pdf>.

[18] Nédellec, Claire. "Machine Learning for Information Extraction." n. page. Print. <http://caderige.imag.fr/Articles/Machine-learning-IE.pdf>.

[19] Ono, Toshihide. "Automated extraction of information on protein–protein interactions from the biological literature." Bioinformatics. 17.2 (2001): 155-161. Print. <http://bioinformatics.oxfordjournals.org/content/17/2/155>.

[20] "PubMed." . US National Library of Medicine/National Institute of Health. Web. 15 Dec 2013. <http://www.ncbi.nlm.nih.gov/pubmed>.

[21] Pustejovsky, J., Castaño, J. M., Zhang, J., Kotecki, M. & Cochran, B. (2002). Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations.. Pacific Symposium on Biocomputing (p./pp. 362-373), .

[22] Yan, Su. "Chemical Name Extraction based on Automatic Training Data Generation and Rich Feature Set."IEEE/ACM Transactions on Computational Biology and Bioinformatics. 1.8 (2012): n. page. Print. <http://www.ncbi.nlm.nih.gov/pubmed/23959634>.